Nucleic Acids Research, 2004, Vol. 32, Web Server issue **W55–W58**
DOI: 10.1093/nar/gkh448

CHOP: visualization of 'wobbling' and isolation of highly conserved regions from aligned DNA sequences

Masato Ohtsuka*, Shohei Horiuchi, Jerzy K. Kulski[1], Minoru Kimura and Hidetoshi Inoko

Division of Basic Molecular Science and Molecular Medicine, School of Medicine, Tokai University, Bohseidai, Isehara, Kanagawa 259–1193, Japan and [1]Centre for Bioinformatics and Biological Computing, School of Information Technology, Murdoch University, Murdoch, 6150 Western Australia, Australia

Received February 11, 2004; Revised and Accepted April 20, 2004

ABSTRACT

The web software CHOP was developed to visualize the 'wobbling' in the third codon position of aligned DNA sequences. The simple features of this tool allow users to easily find regions suspected of containing coding sequences (CDSs). The program also allows visualization of the nucleotide diversity between two genomic or gene sequences by graphically plotting the percentage identity between the two sequences. CHOP can also isolate highly conserved regions within both CDSs and non-CDSs. Highly conserved regions within CDSs include the regions with lower rates of synonymous substitution in which nucleotide sequences are expected to be under strong selective pressure. CHOP is available at http://bunsei2.med.u-tokai.ac.jp:8080/~ohtsuka/cds_finding.html.

INTRODUCTION

As more complete or near-complete genomic sequences of various species continue to become available for molecular and biological analysis, one of the challenges for the biologist is to extract functionally important regions from vast quantities of sequence data. Cross-species genomic comparisons have allowed functionally important regions to be identified within evolutionarily conserved sequences. As a first step in conducting cross-species comparisons, genomic sequences of two or more species need to be aligned. Since order and orientation of functional sites should be conserved, it is preferable to compare between 'non-distantly related' species such as human and mouse, or medaka and *Fugu* [estimated to have branched approximately 75 and 60–80 million years ago, respectively (1,2)]. However, in these comparisons, non-functional regions are still aligned owing to insufficient periods of evolution (1,3). Hence, in 'non-distantly related' species, functional regions such as the coding sequence (CDS) need to be distinguished from the other aligned regions.

The web software 'coding-region hunting online program' (CHOP) was originally developed to find regions with coding potential within aligned data. When aligned sequence data are analyzed at all three codon positions, putative 'wobbling' in the third codon position is visualized in the output graph. We therefore envisioned that this 'wobbling' pattern could be used to identify CDSs. In addition, we added a new function to CHOP in order to detect and isolate highly conserved regions within both CDSs and non-CDSs. Highly conserved regions within CDSs include regions that show lower rates of synonymous substitutions than we might expect from evolutionary distance between species. We have previously identified and reported on a highly conserved region within the 5′ coding region of the *zic4* gene where the conserved nucleotide sequences were expected to be under strong selective pressure (3). CHOP is a helpful tool for further research into such regions.

METHODS

The CHOP online program was developed using the Perl language and it resides on a Linux server machine. As a first step in the CHOP analysis, the aligned sequence is regrouped into three (first, second and third) phases and the percentage identity between the two species is calculated for each phase (Figure 1). The window size is set at 60 bp and advanced by 12 bp (4 bp from each of the three phases). To circumvent the possible frame shifts that may have been caused by errors in nucleotide sequencing, gaps in one of the species (the upper sequence in the alignment) are removed. See the CHOP website for further information on the possible influences of gaps on the CHOP output. The graphs of data, such as the percentage identity of the three phases of the aligned sequences, are then plotted using Gnuplot (http://www.gnuplot.info/).

*To whom correspondence should be addressed. Tel: +81 463 93 1121 (ext.) 2682; Fax: +81 463 96 2892; Email: masato@is.icc.u-tokai.ac.jp

When percentage identity in any one of the three phases is significantly lower than that in the others, probably due to 'wobbling' in the third codon position, vertical brown bars are displayed in the graphical output. Regions with clusters of >5–6 brown bars are considered to contain CDSs. The brown bars are drawn in the program when (i) two of the three phases exhibit >85 and >87% identity and the remaining phase has <80% identity, (ii) two of the three phases exhibit >95 and >90% identity and the remaining phase has <90% identity, or (iii) there is ⩾25% difference between the phases with the lowest and median values. Results can be seen as a diagram [in portable network graphics (PNG) format] and as textual raw data.

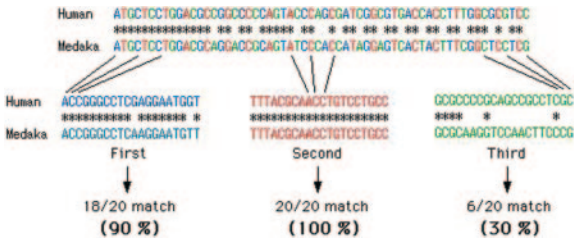

Figure 1. The principle of the CHOP method. When DNA sequences from two species are sorted into three phases according to nucleotide number in the codon, 'wobbling' of the third codon may be seen specifically in the CDS (low rate of similarity in the third phase). The figure shows a sequence comparison between the first 60 bp following the start codon of the *zic1* gene of human (NM_003412) and medaka (included in AB102768).

Table 1. Summary of the comparison of the mutation rate in each codon position of 191 genes (human/mouse comparison)

	% Identity of each codon position			% of regions with brown bar in CHOP output
	First	Second	Third	
Average ± SD	91.8 ± 5.4	94.4 ± 4.5	73.1 ± 6.2	61.4 ± 18.5
Median	93.5	95.8	73.3	65.0
Range	68.6–98.9	73.9–99.5	55.9–91.2	9.2–95.2

The 191 genes located on syntenic chromosomes (human chromosome 14 and mouse chromosome 12) were aligned and mutation rates calculated in each codon position after gaps were manually removed from the alignments. Various percentage identities are detected in different genes. Many factors, including codon bias, can affect the variations in percentage identity. Detailed data (mutation rates in each gene) and a graphical view of the CHOP output are shown on the CHOP website. Percentage identities vary not only in different genes but also in different regions of a gene. Out of 191 genes, 141 show brown bars at >50% of the coding region (see CHOP website).

A summary of the mutation rates calculated by CHOP for each codon position of 191 genes located on the syntenic chromosomes of human (chromosome 14) and mouse (chromosome 12) is shown in Table 1 and on the CHOP website. These results show that the percentage identities of the three codon positions vary not only in different genes, but also in different regions of a gene. Since this may be due to the existence of non-conserved amino acid regions, we also examined the mutation rates of each codon position in both conserved and non-conserved amino acid regions in a comparison between the human and mouse coding sequences (Table 2). As expected, we detected a particular excess of mutations in the third codon positions of conserved amino acid regions. Interestingly, even for non-conserved amino acid regions, 22.9% still generated brown bars. Although mutation rates between different genes or different regions of a gene (Table 1; CHOP website) can vary, possibly due to the influences of various factors such as codon bias, ~85% of the regions that code conserved amino acids generated vertical brown bars in the CHOP output of a comparison between the human and mouse coding sequences (Table 2). In addition, we can obtain the locations of the highly conserved sequences from the aligned data by using CHOP, which extracts and highlights such regions when the identities of all three phases are >90%. The regions of highly conserved sequences are indicated in the outputs by purple bars above the visualized graph (Figure 2A) and/or retrieved as textual data of the sequences and location (Figure 2C and D).

Table 2. Example of a comparison between human and mouse

	Length of aligned sequences (bp)	Average % identity in each codon position (±SD)			% of highly conserved regions with purple bar[a]	% of regions with brown bar
		First	Second	Third		
Coding region[b]						
Conserved amino acid[c]	516 903	98.1 ± 3.2	100.0 ± 0.4	74.5 ± 11.8	10.7 (1.1)	84.7
Non-conserved amino acid[d]	64 734	43.1 ± 12.1	53.2 ± 12.5	57.1 ± 12.4	0.0 (0.0)	22.9
Total	581 637	92.0 ± 9.3	94.8 ± 8.2	72.6 ± 12.5	8.9 (1.1)	61.7
Non-coding region[e]	249 044		63.5 ± 16.4[f]		3.2 (0.9)	4.6

In a human and mouse comparison, 61.7% of the coding region (84.7% of the region that codes conserved amino acids) generates brown bars in the CHOP analysis. Of the region that codes non-conserved amino acids, 22.9% still generates brown bars. This could be due to the substitution of residues with similar properties (such as the change of E to D, and vice versa). Even in the non-coding sequence, 4.6% of the non-coding region showed brown bars. This may be a pseudo-positive signal or unidentified genes.

[a]Numbers in parentheses show the rates of highly conserved regions that exist across >180 bp.

[b]Coding sequences located on syntenic chromosome (human chromosome 14 and mouse chromosome 12) were aligned by BLASTZ. As a result of manually removing the gap region from the aligned sequence, 581 637 bp remained.

[c]Of the aligned sequence 88.9% (516 903 bp) encoded conserved amino acids.

[d]The remaining 11.1% (64 734 bp) coded for non-conserved amino acids.

[e]Non-coding sequence of intergenic region between *plscr-r* and *zic4* genes (3).

[f]For the non-coding region, the average percentage identity was calculated without regrouping into three phases by using another Perl program developed only for this purpose (not shown).

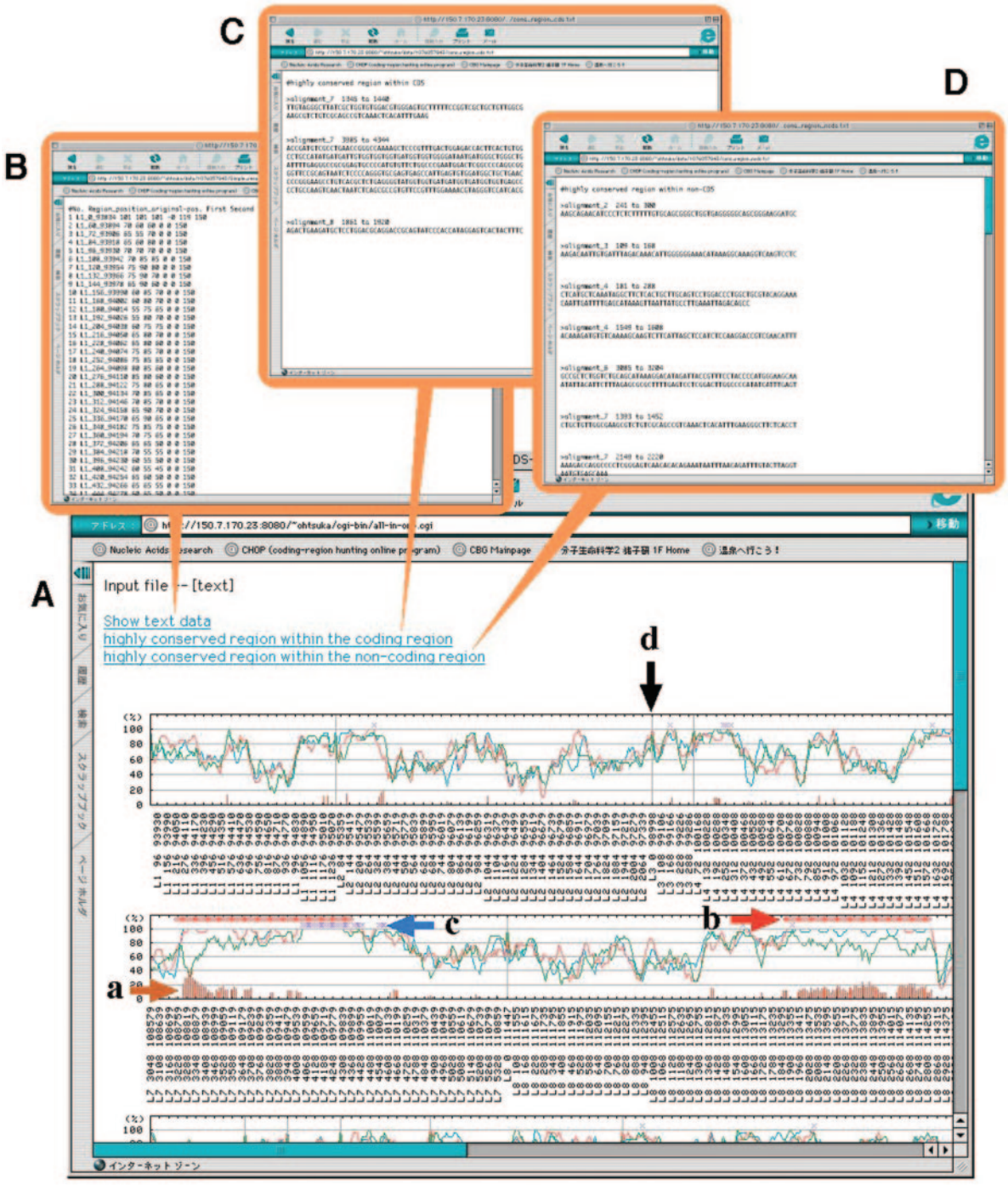

Figure 2. Example of CHOP output. The Medaka genomic region (a part of AB102768) was compared with the corresponding region of *Fugu*. (**A**) The graphical output of CHOP. The percentage identities of three phases are plotted in blue, red and green on the graph. The clusters of >5–6 vertical brown bars (brown arrow a) indicate regions suspected of containing CDSs. Red bars above the graph (red arrow b) are seen in coding regions of known genes. Purple bars above the graph (blue arrow c) indicate highly conserved regions. The local alignments are each separated by vertical gray lines (black arrow d). Letters and numbers on the x-axis indicate the number of the local alignments (e.g. L1, L2), nucleotide number within the local alignments (e.g. 96, 156) and nucleotide number within the genomic region (e.g. 93930, 93990). (**B**) The raw data which is used to build the graph. (**C** and **D**) Highly conserved sequences within CDSs and non-CDSs, respectively.

INPUT

CHOP provides four categories of input form (input forms 1–4). Input form 1 requires a BLASTZ alignment file (traditional textual form of the alignments) as the input file. This can be prepared at the CHOP website or the PipMaker (4) website. The location of known genes within a genomic sequence can be also submitted as an input, and the gene locations will then be indicated in the output as red bars above the graph of the percentage identity plots (Figure 2). This is an optional feature, but it may help to distinguish unknown genes from known genes. In addition, providing the gene locations as part of the input is important to isolate the locations of the highly

conserved regions within a CDS. When input form 1 was used in our system to compare genomic sequences of mouse and human, CHOP ran up to 2.7 Mb of genomic sequence (corresponding to 820 kb of aligned sequence) within 3 min. Input form 2 offers the possibility of uploading two sequences. An alignment is generated by the BLASTZ program included in this web tool and is processed by CHOP. A gene location file can also be included as part of the input. Input form 3 has been made for the analysis of FASTA inputs, where the FASTA alignment file (text) of two sequences is directly processed by CHOP. Input form 4 requires only a single sequence file. It is aligned with the genomes of selected species by the FASTA program and subjected to CHOP analysis. Human, mouse, zebrafish and *Escherichia coli* databases are currently available at our CHOP site and other databases, such as the *Arabidopsis*, yeast and fly, will be added to our site in the near future. For further information regarding the format of the input data, see our web page (http://bunsei2.med.u-tokai.ac.jp:8080/~ohtsuka/cds_finding.html).

OUTPUT

CHOP generates four kinds of output: (i) the plotted graph of the percentage identity of the three phases of the aligned sequence to visualize the putative codon 'wobbling' and highly conserved region; (ii) the raw data which is used to create the graph; (iii) the list of highly conserved DNA sequences within CDSs; and (iv) the list of highly conserved DNA sequences within non-CDSs. Figure 2 shows an example of the CHOP results. In addition, input forms 2 and 4 in the CHOP analysis generate alignments as an output.

CONCLUSIONS

CHOP provides a web-based tool to visualize the 'wobbling' in the third codon position and to isolate highly conserved regions from aligned data. From our assessments, CHOP has worked well for the human/mouse, medaka/*Fugu* and human/medaka DNA sequence comparisons, but not for comparison of closely related species such as human/chimpanzee (see the CHOP website for some examples and further information). We also found that >60% of CDSs were detected in a human/mouse comparison by using CHOP. Although there are sophisticated methods utilizing the codon 'wobbling' or K(A)/K(S) ratio for gene prediction (5–7), the originality of the CHOP program comes from its online accessibility, relative simplicity and easy visualization of the outputs. This user-friendly online program can be used as a complementary method to the other gene finding programs. Extracting highly conserved regions from aligned data is useful for finding functionally important sequences. In cases where the highly conserved region is located within a known CDS, low rates of synonymous substitutions are expected and the nucleotide sequence within the conserved region may be under strong selective pressure, such as codon bias, secondary structure of mRNA, existence of opposite strand transcripts, exonic enhancer and silencing sequences or other unknown factors. To our knowledge, there is no program for detecting and isolating such highly conserved regions. Another useful feature of CHOP is that the first, second and third nucleotides of the codon are plotted visually across the DNA sequences as percentage identities in the graphical output. This provides a helpful visual estimate of the sequence diversity across the genic and/or genomic regions of different species or even different genomic haplotypes within the same species, highlighting interesting regions for further comparative studies. Hence, we believe that the program CHOP will be a helpful tool for teaching bioinformatics and for research on codon 'wobble' and the structure and function of highly conserved genomic regions.

ACKNOWLEDGEMENTS

We are grateful to Dr Tetsushi Yamagata for providing us with an environment for developing this program. We thank Drs Masahiro Sato and Natsuko Kikuchi for proof-reading the manuscript.

REFERENCES

1. Waterston,R.H., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J.F., Agarwal,P., Agarwala,R., Ainscough,R., Alexandersson,M., An,P. *et al*. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
2. Wittbrodt,J., Shima,A. and Schartl,M. (2002) Medaka—a model organism from the Far East. *Nature Rev. Genet*., **3**, 53–64.
3. Ohtsuka,M., Kikuchi,N., Ozato,K., Inoko,H. and Kimura,M. Comparative analysis of a 229 kb medaka genomic region, containing the *zic1* and *zic4* genes, with *Fugu*, human and mouse. *Genomics*, **83**, 1063–1071.
4. Schwartz,S., Zhang,Z., Frazer,K.A., Smit,A., Riemer,C., Bouck,J., Gibbs,R., Hardison,R. and Miller,W. (2000) PipMaker—a web server for aligning two genomic DNA sequences. *Genome Res*., **10**, 577–586.
5. Baillie,D.L. and Rose,A.M. (2000) WABA success: a tool for sequence comparison between large genomes. *Genome Res*., **10**, 1071–1073.
6. Nekrutenko,A., Makova,K.D. and Li,W.H. (2002) The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome Res*., **12**, 198–202.
7. Nekrutenko,A., Chung,W.Y. and Li,W.H. (2003) ETOPE: evolutionary test of predicted exons. *Nucleic Acids Res*., **31**, 3564–3567.